Appendix A

Rapid Spanning Tree Recovery

Table of Contents

1 INTRODUCTION

This document describes the Rapid Spanning Tree Recovery,**RSTR**. This improvement of a Spanning Tree algorithm allows rapid recovery of a spanning tree.

RSTR uses the following concepts:

- **Manual configuration rules** ensure that there is a root switch and a root candidate switch, which can replace the root without changing the topology
- **RSTR core distance** – switches can determine their distance to the RSTR core and use this core distance as the bridge priority thus enforcing certain structure on the spanning tree
- **Upstream port selection**
- **Rapid switchover** - Provided that the root port is upstream, a switch can perform a rapid switchover replacing a broken path to the root with a new path, without significantly changing the topology of the network

Applicable IEEE Documents:

[STD 1]	Std 802.3.x-1997 and 802.3y-1997
[STD 2]	Std 802.3.u-1995
[STD 3]	Std 802.1D

2 GLOSSARY

SPECIAL TERMS

This document uses capitalized port state names. The Spanning Tree port states are prefixed STP_ , e.g. **STP_FORWARDING**. The RSTR port states are prefixed RSTR_ , e.g. **RSTR_ACTIVE**.

ABBREVIATIONS

ASPC	Automatic Switch Priority Configuration
BPDU(A)	Hello message (BPDU) modified with the ASPC flag
BUP	Blocking Upstream Port
FDP	Forwarding Downstream Port
FUP	Forwarding Upstream Port
PDU	Protocol Data Unit, a.k.a. packet
RPG	RSTR Port Group
RSTR	Rapid Spanning Tree Recovery
SMG	Shared Media Group
STG	Spanning Tree Group
STP	Spanning Tree Protocol
TCN	Topology Change Notification
VLAN	Virtual LAN

TERMS

Active RSTR Topology

Core port
Root candidate switch
Rapid switchover
Root switch
RSTR core switch
RSTR mode
RSTR port mode

3 RSTR CONCEPTS AND TERMINOLOGY

This section introduces definitions of switched network elements and concepts, which are used in this document. This section is intended to provide the overview of the algorithms and protocols at the heart of Rapid Spanning Tree Recovery **RSTR**.

The details and analysis of special cases are presented in subsequent chapters.

This section derives **RSTR network design rules**.

3.1 WELL-STRUCTURED CAMPUS NETWORK

Many campus, switched networks are structured as a core high capacity network - typically meshed for redundancy – and a hierarchy of switches.

The Spanning Tree Protocol (abbreviated to **STP**) converts an arbitrary mesh network into a loop-free topology originating at the root switch. Thus, a non-hierarchical network can be automatically converted to a hierarchical network.

Previously, such 'conversion' of arbitrary topology into a tree-topology is based on 'arbitrary' parameters in switches:

- The root of tree is a switch with the lowest MAC address
- As a side effect of this 'random' operation, the root could be a low-capacity switch thus creating a bottleneck in the network.

An additional parameter – path cost - helps to ensure that when selecting between links to the root, the highest capacity links (lowest cost) are preferred over low capacity links. Path cost may be a function of the bandwidth, high bandwidth results in a low cost.

In a RSTR capable network, a network operator can control where the root is placed: a natural place for a root of the tree is one of the highest capacity switches in the core of the network.

The convergence to a loop-free topology can be a time-consuming process. Furthermore, STP does not address the issue of how faults are detected and what action can be taken to optimize the convergence. RSTR provides fast convergence (rapid switchover) after topology change. It avoids the slow STP recalculation, which may take up to 50s. In hierarchical networks RSTR switch knows alternate paths to the STP root and it allows performing a rapid switchover operation. This switchover requires special operation of RSTR switches, and it requires that a few network design rules be followed to prevent single failures to trigger the slow spanning tree recalculation.

A typical hierarchical network design, which includes 3 layers, is shown below.

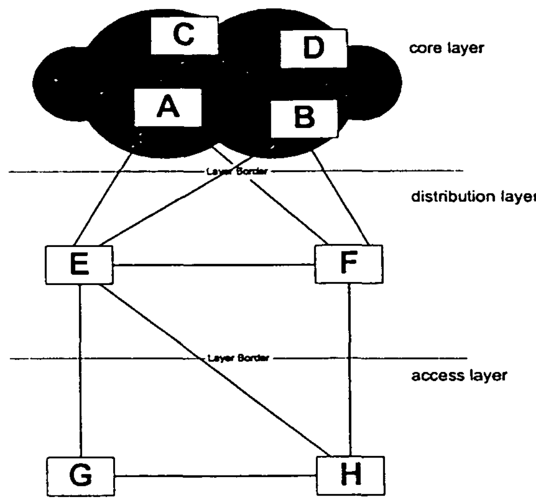


Figure 3-1 Hierarchical Switched Network

RSTR focuses on a **fast recovery of the spanning tree in switched networks**, i.e. networks where inter-switch links are dedicated, point-to-point links. RSTR includes, however, provisions for shared media and ATM.

RSTR ensures that the spanning tree topology follows this intended network hierarchy:

- The root is selected among top (core) switches
- The tree propagates downward towards the branches which are at the edge of the network

Furthermore, RSTR ensures that any RSTR enabled switch knows its topological position, whether the ports connect a switch to upper or lower layer switches and allows this information to be utilized for rapid switchover from a failing forwarding port to a backup port.

3.2 CORE DISTANCE

A fundamental concept in RSTR is the **core distance** and automatic learning of STP Bridge priorities. The network core is, preferably, a set of switches in an arbitrary meshed topology, which create the top layer of hierarchical network and are the STP root candidates. To enable a fast root recovery, there should be two RSTR core switches, and for the purpose of fast failure detection, they should be interconnected by a direct link.

Switches may be manually configured with information identifying whether they are core or non-core switches. This will set RSTR core switches' STP Bridge priority to 0, thus making them top candidates to become root switches.

> *The* **network RSTR core** *consists of the root switch and its directly connected root candidates.*

RSTR capable switches may, by default, be enabled for the RSTR operation. Those switches will learn their priority from the Spanning Tree Protocol Hello PDUs, i.e. BPDUs.

This automatic learning process is a recursive process starting from the root candidates. It quickly stabilizes and it enables switches to learn their core distance:

> ***RSTR core distance*** *is the shortest path from the switch to the network RSTR core, measured in hops. Each network layer adds one hop to the path. The core distance is counted from 0.*

The learned core distance is used in the STP bridge priority thus influencing the root selection.

This automatic process is called **Automatic Switch Priority Configuration ASPC**.

> *All switches that have the same core distance are descibed, herein, as belonging to the same* **network layer**.

After having determined their core distance, and thereby their STP Bridge priorities, switches can determine whether each of their ports is connected to an upstream (upper layer), downstream or peer switch. This information is in turn used to create recovery groups, which allow a switch to perform a rapid switchover in case the forwarding root port fails.

3.3 CONFIGURATION OPTIONS

A switch can be enabled to operate in the **RSTR mode**. If the RSTR mode is enabled, then, in general, all switch ports operate in the RSTR port mode.

Similarly, if a switch operates in the RSTR mode, a port can be configured to operate in a standard mode, i.e. with RSTR disabled on this mode. Such mode of port's operation can be selected for the reasons including the following:

- An RSTR switch connects through this port to a part of the network where RSTR is not operating
- An RSTR switch connects through this port to a switch where interoperability is in question

A port operating in the RSTR port mode can send modified BPDUs. The modified BPDUs may be similar to standard (IEEE) Spanning Tree Protocol Hello (e.g., BPDU) messages, modified by a single bit in the flag in the options field. Therefore RSTR BPDUs use a different (non-IEEE standard) protocol version number. For compatibility with standard BPDUs, every second RSTR BPDU is modified so that non-RSTR switches will continue to receive standard BPDUs. This may improve interoperability between RSTR and non-RSTR switches.

A switch in the RSTR mode has RSTR "components" (e.g., circuitry to implement RSTR algorithms and protocols) active. An RSTR switch may automatically learn some parameters, like STP priority, and make faster STP convergence. When switch's RSTR mode is disabled, the automatically configured RSTR parameters may be deleted and the switch may thereafter function in a standard way.

The **RSTR core switch mode** may also be manually configured. When RSTR switch core mode is enabled, the switch uses STP Bridge priority of 0 thus making it top candidate to become a root of the spanning tree.

> ***RSTR Network Design Recommendation 1:*** *In a RSTR network there are (preferably) at least two switches that are manually configured as RSTR core switches. Those switches should be connected by a direct link.*

This leads to the following manual configuration options:

- RSTR mode, default enabled
- RSTR port mode, default enabled when RSTR mode enabled
- RSTR Core switch mode, default disabled

3.4 ACTIVE RSTR TOPOLOGY

The set of RSTR enabled switches and their RSTR enabled ports constitute the RSTR topology. This topology is reduced to the active RSTR topology, which consists of RSTR switches and their **RSTR active ports.**

The RSTR active ports are RSTR enabled ports, which connect to another RSTR enabled ports. Since an automatic detection of such ports is recommended – because of network changes, manual reconfigurations, etc. – a special convergence protocol overlaid on top of standard Spanning Tree Protocol is used. This protocol uses a special flag in Hello Messages (BPDUs), which signals to the neighbor switch that this port is RSTR enabled. This protocol allows two connected ports to auto-negotiate the correct mode of operation. Those special BPDUs are called **BPDU(A)** – to indicate the use of A-flag.

This handshake signal may be sent in every second BPDU on RSTR enabled ports thus enabling a variety of possible network reconfigurations, e.g. manual switch and port reconfigurations, introductions of new switches into the network, etc.

> *The* ***active RSTR topology*** *consists of RSTR enabled switches and RSTR enabled ports connected via links to other RSTR enabled ports.*

3.5 AUTOMATIC SWITCH PRIORITY CONFIGURATION

To determine the core distance and hereby their STP Bridge priority the switches observe neighbor switch's STP Bridge priority received in BPDU(A) frames. The RSTR core switches can set their STP Bridge priority to 0 to ensure that they win the root election process. Layer 1 switches will observe this priority in BPDU(A) frames and conclude that they are at layer 1.

Actually, the RSTP Bridge Priority Selection Algorithm uses neighbor switches' RSTP Bridge priority in BPDU(A) observed on all RSTR active ports, finds the minimum value of those observed values and uses this value incremented by 1 as its RSTP Bridge priority.

This automatic process is called Automatic Switch Priority Configuration**ASPC**. The STP Bridge priorities are thus quickly learned, are changed on the flight, and result in a proper Spanning Tree topology, which is 'aligned' with the intended network hierarchy and follows the natural bandwidth hierarchy.

The result of ASPC is that the active network topology is well structured and that all RSTR switches know their STP Bridge priority.

By continuously observing received RSTP Bridge priorities, and deriving its own RSTP Bridge priority a switch can determine whether there are any topology changes. Furthermore, a switch can determine for each RSTR active port whether the port connects to upper or lower network layer.

3.6 PORT ROLES

For each RSTR active ports an RSTR enabled switch can determine a**port's role**. This port's role is determined by comparing the STP Bridge priorities observed on a port with the derived own STP Bridge priority. The port's role can be:

- **Upstream Port,** falling into **Forwarding Upstream Port** or **Blocking Upstream Port (FUP, BUP)** – connecting to an upper layer switch.
- **Downstream Port DP** – connecting to a lower layer switch.
- **Peer Port PP** – connecting to a switch in the same network layer.

- **Core Port** – connection between two RSTR core switches. This is a special case of peer link, which connects core switches. The rules for handling the STP Bridge priority for such a port are different from other ports.

Additionally, we define one more concept:

- **Backup Port BP** – A switch may see an alternative path to the root through a backup port.

3.7 RSTR PORT GROUP

The port role makes it possible to create **RSTR Port Groups RPG** consisting of the root port and the backup ports. RSTR can also take advantage of peer links: peer ports can become members of RSTR groups and can be used for the rapid switchover. This increases the probability of successful recovery. Thus, the definition of RPG includes the peer ports:

> *The **RSTR Port Group** consists of the upstream or peer forwarding port (FUP or FPP) which is the root port and backup ports which are blocking upstream ports BUP and blocking peer ports BPP. The blocking ports provide an alternate path in case the currently forwarding port fails.*

The backup ports can be used to replace the root port if this port (or the neighbor switch) fails.

The configuration of the path cost is important. If the best cost towards the root is in the downstream direction the root port may become downstream. In this case the sending of BPDU(A)s will be disabled to prevent downstream switches from including their upstream links towards this switch in their RSTR Port Group. This effectively excludes a possibility of creating forwarding loops. The network will function properly but a part of it will be excluded from the rapid switchover mechanism.

Thus, the following procedures are recommendation:

> ***RSTR Network Design Recommendation 2:*** *The network should be configured so the direction of the least cost towards the root matches the direction of increasing priority.*

In the degenerated case, the RSTR port group contains only FUP. In such a case, there are no backup ports and no fast recovery is possible. Therefore another RSTR network design rule defines an additional requirement to enable fast recovery:

> ***RSTR Network Design Recommendation 3:*** *Its is recommended that each non-RSTR core switch connects to two upper layer switches.*

Such a partial design is shown below.

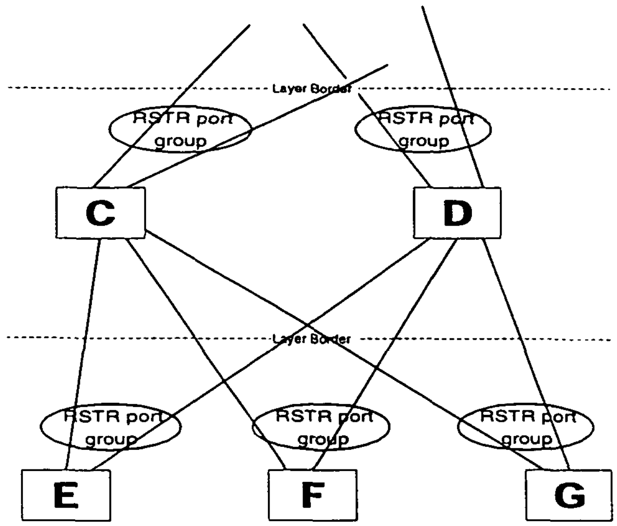


Figure 3-2 Redundancy Requirements and RSTR Port Groups

This design rule ensures that RSTR Port Groups contain at least 2 ports, one STP_FORWARDING and one STP_BLOCKING. RSTR can operate even when this design rule is not followed, but there will be cases where a single link/port/switch failure will de-stabilize the spanning tree.

RPGs are dynamically maintained on the basis of port roles. Those port roles will change if the active network topology changes.

Adding a new member port to RPG requires caution. Since the port addition operation is not synchronized between two neighbor switches, the switch which adds a new port to RPG should wait some time before it can use this port for a switchover. This is done in order to ensure that the neighbor switch has changed the port's STP state to STP_FORWARDING:

> ***Adding new member to an RSTR Port Group:*** *When a switch adds a new member port to RPG, it should wait for a recommended period of twice the STP Forward Delay time plus one second to prevent race conditions before the new port is used as a backup port. This port's RSTR state is called* ***RSTR_BLOCKING_TIMING*** *when waiting to become a full member of RPG. A port, which immediately can be used as a backup port, is called* ***RSTR_BLOCKING_STANDBY.***

RSTR rapid switchover operation may be performed between ports in the same RPG. The switch may have as many RSTR Port Groups as there are Spanning Tree instances in the switch.

Note that ports, which belong to the same RSTR Port Group, may have different costs to the root.

3.8 RAPID SWITCHOVER

The presence of a backup port in the RSTR port group makes it possible to perform the rapid switchover:

> *If the root port fails, then RSTR can, depending on various conditions, perform a rapid switchover to a backup port.*

The exact conditions for performing the rapid switchover are described in the detailed specification sections. Such switchover makes only a local change to the active topology, and is 'invisible' to the rest of the network.

To take advantage of the new path towards the root, the switch performing the switchover should immediately age out its forwarding database.

The switch performing the switchover should send the Topology Change Notification (TCN) frame. RSTR enabled switches en route to the root receiving TCN-frames should immediately age out their forwarding databases.

The upper layer switch, which is disconnected from the switch performing the switchover, should also send TCN-frame. This frame is sent on its root port, thus immediately aging out forwarding databases in all switches along the old, broken path to the root. ***[clarification: standard STP generates also TCN frames but requires all switches to use short aging timer, typically 15 seconds – during this time forwarding paths may be wrong]***

Those requirements are listed below:

Topology Change Notification and Address Aging Requirements:

- An RSTR switch performing the rapid switchover should send a Topology Change TCN-frame on its new root port and age out its forwarding database
- An RSTR switch which detects a failure of its forwarding downstream port should send a TCN-frame on its root port
- An RSTR switch receiving a TCN-frame should immediately age out its forwarding database

Those requirements are needed to conform to the 3 seconds recovery time for ClearSession.

The figure below shows an example of a link failure and TCN-frames sent by switches F and H. Alternatively, if switch H have failed, then both F and G would have sent TCN frames.

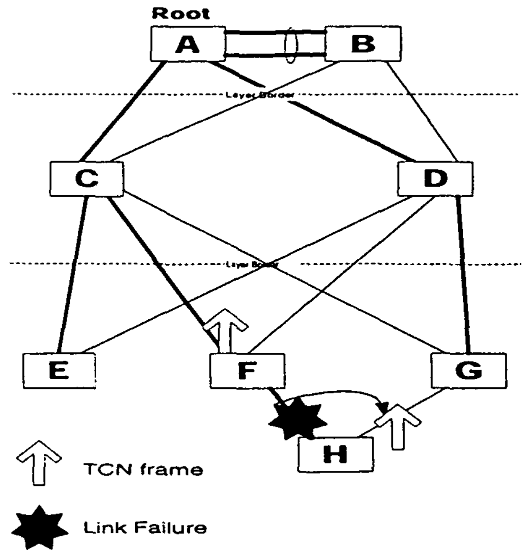


Figure 3-3 Rapid switchover and Topology Change Notification

3.9 ROOT RECOVERY

To maintain a stable spanning tree , the root should always be present and operational. To help ensure this, at least two RSTR core switches with a direct link between them may be used. The use of the STP Bridge priority of 0 ensures that one of those RSTR core switches becomes the**root switch**, while the others become the **root candidate switch.**

The root candidate switches monitor the presence of the root switch, and can immediately take over (or rather attempt to take over) if the root fails. A root candidate will assume that the root "died" by missing BPDUs on its core port.

If the direct link between core switches fails, and the root is still operational, then the network topology will be significantly changed as shown in the figure below where the root candidate's core distance becomes 2 after the link failure. Such a change of the network topology caused by single failure results in a de-stabilization of the spanning tree (in a part of the network).

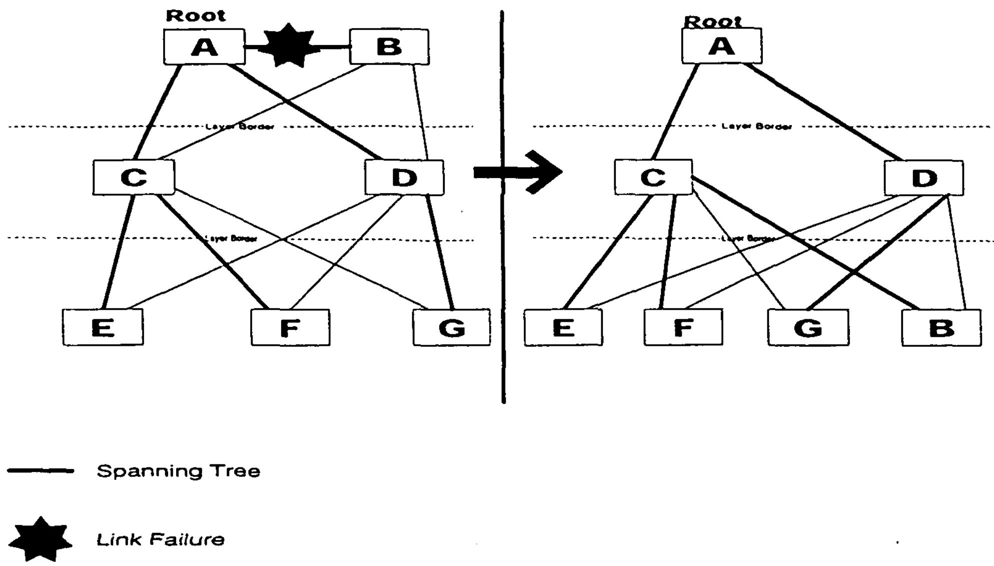


Figure 3-4 RSTR Core Link Failure – Drastic Topology Change

To prevent such single failures to impact recovery times RSTR, the following RSTR network design rule is recommended:

> **RSTR Network Design Recommendation 4:** *It is recommended that the direct link between RSTR core switches be an aggregated link consisting of at least two links. (an aggregated link is dual-link or triple-link interconnection in which the links perform load sharing and can withstand the failure of an individual link).*

The foregoing design rule also helps to ensure spanning tree stability in cases of single-link failure.

The figure below shows an outline of 3-layer network designed according to the RSTR design rules.

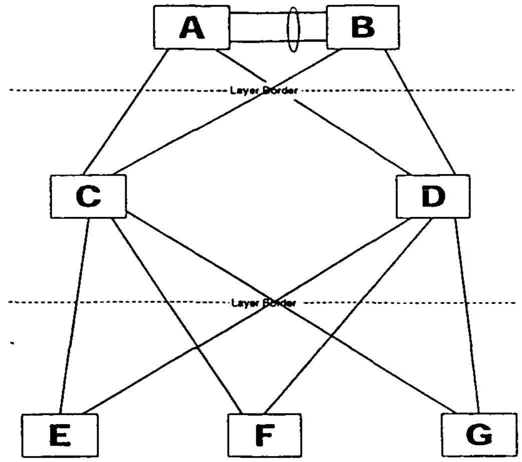


Figure 3-5 Fully Redundant Switched RSTR Network

The figure below shows a more detailed example with most of the concept described so far.

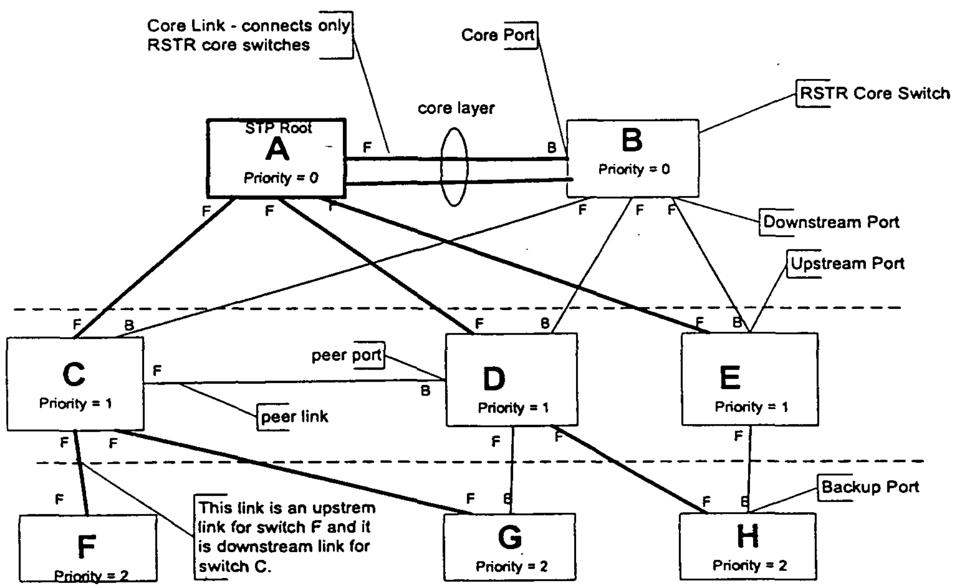


Figure 3-6 Well Structured Multi-Layer Hierarchical Switched Network

Figure 3-6 shows well-structured multi-later hierarchical switched network. This figure shows the following definitions:

- Core Layer, Distribution Layer and Access Layer.
- Downstream, Upstream, Core and Peer Ports.

- Downstream, Upstream Switches.
- Downstream, Upstream and Peer Links.

3.10 FAILURE DETECTION

RSTR switches should perform fast detection of link and switch failures under a variety of conditions, such as:

- **Physical link failures** – if possible, physical layer should be detected and reported to switch RSTR functionality. Detected link failures may trigger rapidswitchover.
  - **Link failures** – A RSTR switch may employ special fast polling to detect failures on links.
- **Switch failures**
  - Neighbor switches may detect switch reboots as link failures.
  - A RSTR switch may employ a hardware watchdog circuitry to reset a switch due, for example, to software failure. The hardware watchdog may be a timer circuit that is reset periodically by the switch software (say, within a 1.5 second period). If the hardware watch dog is not set within the designated period, it may trigger a switch re-boot process to reload switch software.
- **Switch misbehavior:**
  A downstream switch may use missing BPUDs (and BPUD(A)s) on upstream ports as an indication of a switch failure, if the physical layer is still operational. In particular, the root port requires special attention: A downstream switch, which fails to receive BPDU within a designated interval, say 2.0 seconds, may assume that an upper layer switch has failed. The downstream switch may thereafter perform rapid switchover replacing its root port with a backup port. Subsequently, it may generate TCN frame on the new root port. As a precaution, the switch can also turn off, and then turn on its physical layer on the former root port. Should the formerly used upper layer switch still be operational, it can use this physical layer change as a signal to generate the TCN frame.

3.11 SHARED MEDIA CONSIDERATIONS

The shared-media Ethernet segments are supported insofar they are leaf segments, i.e. having no sub-layers.

Dual cables protect the shared media: A single switch can connect by two cables to the same shared-media segment. For improved reliability a stack with two switches can be used, each cable being allocated to different switch.

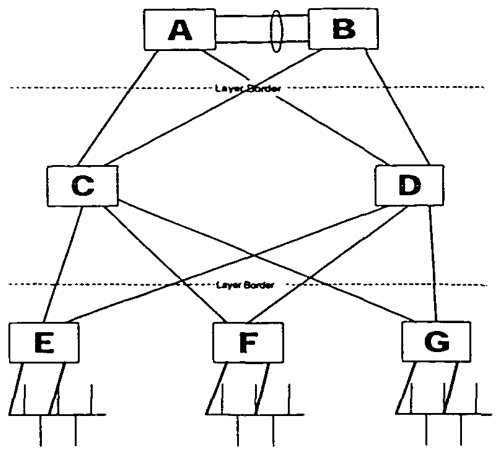


Figure 3-7 Shared Media Design – Shared Media at the Access Layer

Those considerations lead us to the following network design recommendation:

> ***RSTR Network Design Recommendation 5:*** *It is recommended that shared-media segments be at the access layer. To enable RSTR on those segments it is recommended that there are two links from the access switch to the shared media. To protect against single component failures it is recommended to use a switch stack with each link connecting to a different switch within a stack.*

A switch may be able to detect that two ports connect to the same shared-media. Subsequently a switch can create a **Shared-Media Group SMG** and utilize its knowledge of received BPDUs to detect failures and automatically recover from failures by letting the standby port resume transmission of BPDUs which previously have been transmitted by the failed port.

Note, the access layer can extend downwards in the tree (away from the root) and still enable RSTR protection. In the situation shown in the figure below the legacy Ethernet network extending from and below the access layer is still protected by the shared-media group.

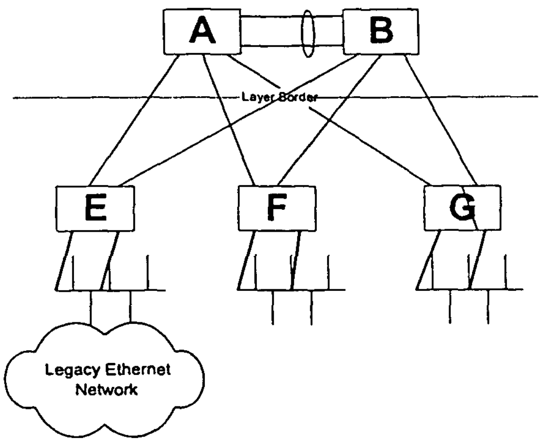


Figure 3-8 Shared Media Design – Protecting Legacy Network at the Access Layer

It is also worth noticing that if a network configuration has two shared media groups in two different switch stacks on the same shared media, then RSTR is still working and protects against failures but obviously it uses 2 stacks of at least 2 switches and 4 links. Such protection works but may not provide additional reliability since each shared media group operates independently of each other. RSTR for shared media generally does not use distributed switches on the same shared-media segment.

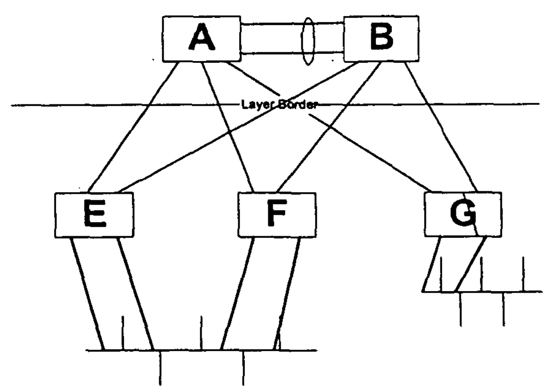


Figure 3-9 Shared Media Design – 2 RSTR Switches at the Access Layer

3.12 SUMMARY OF RSTR NETWORK DESIGN RECOMMENDATIONS

- **RSTR Network Design Recommendation 1:** There should be at least two switches, which are manually configured as RSTR core switches. Those switches should be connected via a direct link.
- **RSTR Network Design Recommendation 2:** The network should be configured so the direction of least cost towards the root matches the direction of increasing priority.
- **RSTR Network Design Recommendation 3:** Each non-RSTR core switch should connect to two upper layer switches.
- **RSTR Network Design Recommendation 4:** The direct link between RSTR core switches should be an aggregated link consisting of at least two links.
- **RSTR Network Design Recommendation 5:** The shared-media segments should be at the access layer. To enable RSTR on those segments, there should be two links from the access switch to the shared media. To protect against single component failures, the use a switch stack with each link connecting to a different switch within a stack is recommended.

4 AUTOMATIC RSTR CONFIGURATION

An RSTR network contains one or more core and a number of non-core switches interconnected by dedicated point-to-point links. RSTR core switches are the STP root candidates. A switch may be manually configured to designate it as a RSTR core switch.

For all other switches, the STP Bridge priority is a function of the distance (measured in number of hops) to the core. ASPC makes sure that the STP Bridge priority is equal to the core distance. For RSTR core switches the distance to the core is 0. Once core switches are defined, the automatic STP priority configuration can begin. All other switches learn their new STP Bridge priority in a recursive way starting from the core switches. The STP Bridge priority is incremented for each switch traversed. For example, the RSTR core switches send priority 0 to its neighbors, its neighbors send priority 1, etc.

The ASPC protocol will force the same priority on all switches in the same layer in the hierarchical network. The STP Bridge priority learnt automatically via ASPC determines the core distance.

The ASPC protocol is a basis for the following algorithms:

- Automatic upstream, peer, and downstream **port role** selection
- Automatic **RSTR Port Group** creation

4.1 RSTR CORE SWITCHES

Core switches can be manually configured to become STR core switches. The STP Bridge priority becomes then 0. Both core switches should be configured to have the same priority, i.e. both the root and the root candidate switch will have the core distance of 0.

The STP Bridge priority is advertised through downstream ports via the STP BPDUs. Those BPDUs have a special ASPC flag set in the *flag* field. The ASPC flag (combined with the advertised root priority) is a key mechanism in the priority learning process.

The STP Bridge priority 0 for core switches can automatically be changed if the direct link between core switches fails as described below.

If a root candidate switch loses its direct link to the root switch, then the network topology has undergone a significant change. This change should be reflected in the STP Bridge priority for the root candidate:

- The root candidate will initially assume that it is a new root, and will send BPDUs based on this assumption

- If the root candidate is not contradicted by a preferred root candidate BPDUs, then it will become the STP root (and it will continue using the STP Bridge priority 0)

- If the root candidate is contradicted by a preferred root candidate BPDUs (from the original or new root) with a message age less than the time passed since the link was lost, then its core distance to the existing root has changed (or a new root has appeared changing the topology). In this case the root candidate should (perhaps temporarily) derive its STP Bridge priority from BPDU(A). This may result in a reconfiguration of the spanning tree. However, such reconfiguration is a result of double failure (aggregated link between RSTR core switches failure) and is not recovered by RSTR.

- If the root candidate is contradicted by a preferred root candidate BPDUs from the original root with a message age not less than the time passed since the root link was lost, then the existing root has failed but the layer 1 switch did not update its root (yet). In this case it should disregard such BPDUs and continue to assume that it is the root.

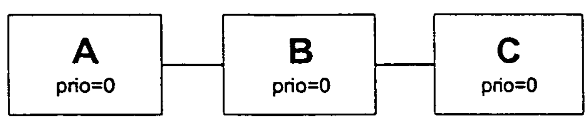
- When the root candidate has re-established its direct link to the root it should revert to using the STP Bridge priority of 0. This will result in a reconfiguration of the spanning tree.

4.2 MULTIPLE RSTR CORE SWITCHES

Proper RSTR configuration may require no more than a designated limited number of priority 0 switches. In the implementation described herein, the maximum number of priority 0 switches may be two -- the root and the root candidate. If 3 or more switches have been configured with priority 0, incorrect configuration may occur.

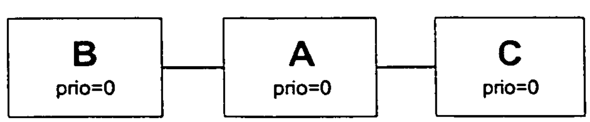
For example, consider the following two cases:

Case 1: The root switch (A) has a direct connection to only one other priority zero switch (B)



In this illustration, A is the root switch. In this situation standard RSTR behavior may force switches with no direct connection to the root to change their priority. Thus a normal RSTR configuration will ensue. In the illustration above, C will automatically move to priority 1.

Case 2: Two or more priority zero switches (B and C) have a direct connection to the root (A)



In this illustration, A is the root switch. This configuration should result in proper operation.

4.3 NON-CORE SWITCHES

The STP Bridge priority is automatically learned on non-core switches by the ASPC protocol.

When RSTR core switches are configured they start to transmit BPDU(A) as defined by BPDU (A) rules on all ports where RSTR operation is enabled. A non-core switch will observe priorities from incoming BPDUs on all ports. The priority is taken from the Bridge ID field in BPDU. This Bridge ID also determines which switch that has sent this BPDU.

The STP Bridge priority is learnt from BPDUs that have the ASPC flag set, i.e. from BPDU(A)s.

If a switch has a manually configured STP bridge priority and the switch operates in the STR mode, then this manually configured priority is replaced by the automatically learned STP Bridge priority. The automatically learned STP Bridge priority is used by the STP function.

If a switch has no manually configured STP Bridge priority, then the default STP Bridge priority is also replaced by the automatically learned STP Bridge priority.

A non-RSTR core switch continuously learns its core distance by observing BPDU(A)s. It selects the lowest core distance observed, increments it by 1, and uses this value as its RSTP Bridge priority.

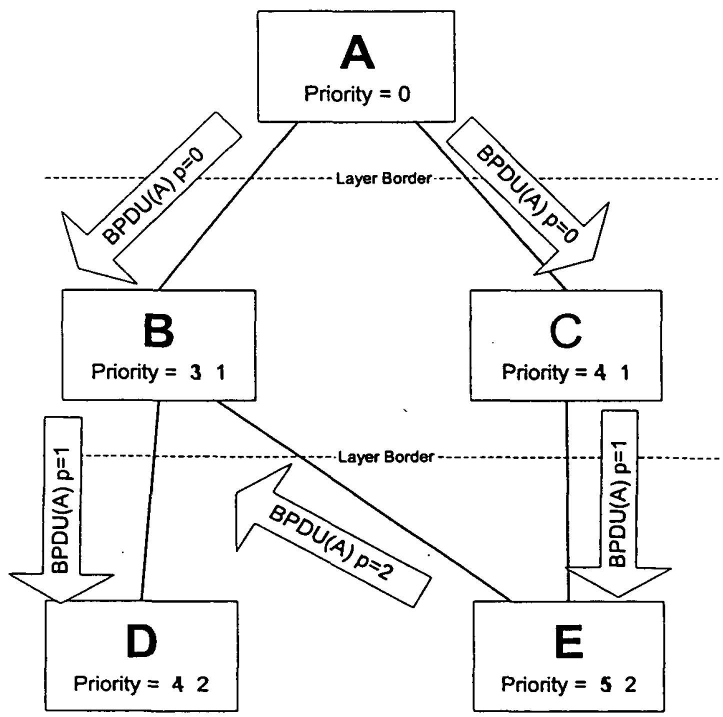


Figure 4-1 Automatic Switch Priority Learning

The figure above shows the concept of automatic switch priority learning, where the lowest core distance received in any BPDU(A) determines the actual core distance and hereby the STP Bridge priority.

4.4 RSTR PORT STATES AND BPDU RULES

The RSTR enabled ports are ports which send (and receive) BPDU(A). Only those ports are considered for the STP Bridge priority learning.

For a switch where RSTR is globally enabled, all ports are default RSTR enabled. The operator may though disable RSTR on a port.

RSTR enabled ports maintain RSTR state which can be:

- **RSTR_CANDIDATE** – the port is a candidate for inclusion into active RSTR topology
- **RSTR_ACTIVE** – the port is included in the active RSTR topology
- **RSTR_SUSPENDED** – the port is receiving conflicting information about the neighbor switch and is excluded from the active RSTR topology

For completeness, there also may be a **RSTR_INACTIVE** state used to indicate that the physical layer is not operational, and a RSTR_DISABLED state to indicate that RSTR or STP is disabled on this port.

The changes between the RSTR status occurs are intimately related to the BPDU(A) forwarding and processing rules.

The rules for the BPDU(A) handling should ensure that:

- Switches automatically learn priorities
- Switches automatically detect on which STP_BLOCKING or Root ports they are connected to RSTR enabled switches & RSTR enabled ports
- For a network, which separates into disjoint networks, each isolated sub-tree can still operate the ASPC algorithm.

The BPDU handling rules and the associated changes between RSTR port state changes for RSTR enabled switches are as follows:

- A switch initially assumes that it is a root and all its ports are downstream ports.
- Every RSTR enabled port is initially assumed to be an **RSTR_INACTIVE.** It becomes **RSTR_CANDIDATE** when the physical layer is operational. For example, when Fast Ethernet auto-negotiation has established the link, or the link is presumed operational.
- The failures of the physical layer change the port's state to RSTR_INACTIVE.
- An **RSTR_CANDIDATE** port observes the following rules:
    - A switch transmits every second BPDU as BPDU(A) on RSTR_CANDIDATE ports and should start with BPDU(A).

- A switch receiving BPDU(A) on a RSTR_CANDIDATE port, will change the port RSTR status to RSTR_ACTIVE and include this port in RSTR algorithms.

- An **RSTR_ACTIVE** port observes the following rules:
  - A switch transmits every second BPDU as BPDU(A) on RSTR_ACTIVE ports.
  - When a switch does not receive BPDU(A) on an RSTR_ACTIVE port for 3 BPDU(A) cycles (default 12 seconds) it changes the port's RSTR state to RSTR_CANDIDATE.
  - When a switch does not receive any BPDU for twice the Hello time (4 seconds) on an RSTR_ACTIVE port, the port state is changed to RSTR_CANDIDATE.
  - If the received standard BPDU from the neighbor switch includes a different STP Bridge priority then previously observed in BPDUs and BPDU(A)s on this port, then the port's state is changed to RSTR_CANDIDATE – this is a safety rule.
- A switch, which receives BPDUs or BPDU(A)s from two or more different switches on the same port, should temporarily exclude this port from the RSTR operation. This could happen, e.g. when an 'intermediate' switch has disabled its spanning tree, due to a mis-configuration, or a transient condition. If the affected port is in the RSTR_ACTIVE state or in RSTR_CANDIDATE state, then it should be moved to the RSTR_SUSPENDED state.
- The RSTR_SUSPENDED state is entered when a port receives a BPDU or BPDU(A) which indicates a different switch than in the previously received BPDU or BPDU(A).
- An **RSTR_SUSPENDED** port observes the following rules:
  - An RSTR_SUSPENDED port observes received BPDUs and BPDU(A)s, and uses an RSTR suspended timer to detect when there is only one neighbor switch sending BPDUs/BPDU(A)s. The RSTR suspended timer is three times the Hello timer, i.e. it has a value of 6 seconds.
  - If the port observes BPDU and BPDU(A) frames from only one switch for the duration of the RSTR suspended timer, the port's state is changed to RSTR_CANDIDATE.
  - BPDUs and BPDU(A)s should still be transmitted on an RSTR_SUSPENDED port, to enable automatic recovery from the improper configuration.
- All BPDUs and BPDU(A)s transmitted should include the STP Bridge priority derived by the ASPC algorithm rules defined in the next section

Note, that standard BPDUs with different STP Bridge priority (than previously observed) result in RSTR_ACTIVE port becoming RSTR_CANDIDATE whereas only BPDU(A)s with different STP Bridge priority force the actual re-calculation of switch's own STP Bridge Priority.

Only designated ports transmit BPDUs. Hence, a designated port will be RSTR_CANDIDATE and only the other end of the link will be RSTR_ACTIVE. In a preferred implementation, RSTR_ACTIVE ports will not send BPDUs (or BPDU(A)s).

The STP Bridge Priority Selection Algorithm's activation and de-activation rules are defined as follows:

- A switch should activate its STP Bridge Priority Selection Algorithm as soon as a single port is in the RSTR_ACTIVE state
- A switch should stop its ASPC algorithm when one of the following condition occurs:
    - RSTR is disabled
    - No port is in the RSTR_ACTIVE state

The BPDU(A) rules ensure a **convergence of the active RSTR topology**, helping to ensure that the following goals are met:

- RSTR switches will eventually force neighbor switches' ports to RSTR_ACTIVE state, if the neighbor switches are configured for RSTR operation
- RSTR switches will not harm non-RSTR switches even if the RSTR is enabled on ports connecting to non-RSTR switches. Ultimately, RSTR can be disabled on ports connecting to non-RSTR switches
- A Manual reconfiguration of RSTR switches and their ports will eventually propagate to neighbor switches
- A total failure of RSTR core switches will still make the remaining RSTR network operate in a proper way

The RSTR enabled port's state follow the state-event machine outlined below:

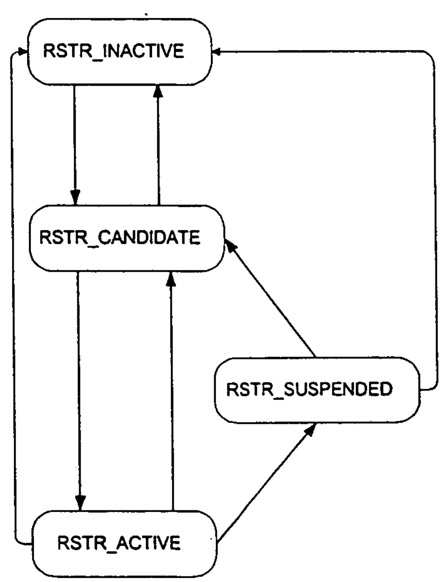


Figure 4-2 RSTR Port State-Event Machine

A port can change to RSTR_INACTIVE as a result of several events:

- Operator controlled disabling of the port
- Operator controlled disabling of STP on the port
- Operator controlled disabling of RSTR on the port
- Failure of the physical layer – link failure
- Port's STP state is STP_LEARNING or STP_LISTENING

4.5 RAPID SWITCHOVER AND TOPOLOGY CHANGE NOTIFICATIONS (TCN)

The topology changes resulting from the rapid switchover can trigger Topology Change Notification (TCN) frames. Those notification frames can be propagated along the new path to the root and along the broken path to the root. RSTR switches should react quickly to these notification frames. This specification recommends that, in an RSTR switch, addresses learnt on the root port and the port on which a TCN frame is received by purged from the forwarding database upon receipt of a TCN frame. This aging procedure may help to achieve rapid recovery, though this may occur at the expense of network flooding.

The recommended TCN processing rules are as follows:

- If a switch performs the spanning tree recovery by a rapid switchover from a failing forwarding port to a backup port, the switch should send TCN-frame on its new root port.
- When a switch detects a failure of its FDP, it should send a TCN-frame on its root port.
  - This switch may be disconnected from the switch performing fast switchover, i.e. it is on the ingress of the broken path. Therefore it should inform switches along the broken path to the root about topology change so they can age out their forwarding databases.
- An RSTR switch receiving a TCN-frame should purge from its forwarding database addresses learnt on the root port and the port on which the TCN frame is received.
- TC-flag and TCA-flag handling follows the standard STP rules.

Note that a switch may receive a TCN-frame due to normal STP processing in the neighbor switch. RSTR switches should purge from their forwarding databases addresses learnt on the root port and the port on which the TCN frame is received.

If a network layer consists of non-RSTR switches, then those switches will still use the standard fast aging timer (15 seconds) instead of immediately aging out the tables.

4.6 STP BRIDGE PRIORITY SELECTION ALGORITHM

The calculation of the STP Bridge priority uses the following rules:

- The algorithm considers STP Bridge priorities derived from Bridge Id's from BPDU(A) on RSTR_ACTIVE ports

- Only STP Bridge Priorities from valid BPDU(A)s should be used, valid BPDU(A)s being
  - They are associated with RSTR_ACTIVE ports
  - BPDU(A) is not invalidated by a better root (as received in another BPDU)
- The learned STP Bridge Priority is calculated as

  MIN { VALID BPDU(A)[NEIGHBOR STP_BRIDGE_PRIORITY] } + 1

The algorithm is driven by:

- Changes in neighbor's STP Bridge priority observed in BPDU(A)s
- Swithover from a forwarding port to a backup port
- Changes in RSTR port states – RSTR_CANDIDATE to RSTR_ACTIVE, or RSTR_ACTIVE to RSTR_CANDIDATE or RSTR_ACTIVE to RSTR_SUSPENDED
- If there are no valid BPDU(A)s – either because there are no RSTR_ACTIVE ports, or none BPDU(A) has been received – the switch should not change its priority.

The last rule helps to ensure that, if a sub-tree is isolated, the top-level switch will become a root according to the standard STP operation.

When the STP Bridge priority is changed it is important that the STP Bridge Priority Selection Algorithm and STP algorithm operate in a proper sequence:

- Received BPDUs and BPDU(A)s should first be processed by ASPC, possibly triggering the STP Bridge Priority Selection Algorithm, and potentially changing the STP Bridge priority
- If the STP Bridge priority has changed, the standard STP algorithm should be activated
- Finally, the received BPDU or BPDU(A) should be passed on to the standard STP algorithm

4.7 BPDU(A) - BPDU MODIFICATION

The IEEE 802.1D standard permits usage of the flag field in the STP Hello BPDU. However, a new Protocol Version Identifier should be used when the flag field is used.

7	6	5	4	3	2	1	0
TCA	0	0	0	0	0	ASPC	TC

Figure 4-3 BPDU(A) Flag Field

> *This specification defines the Protocol Version identifier value of* ***1000 0001B*** *(i.e.* ***0x81*** *hexadecimal) to be used for BPDU(A).*

In general, only RSTR capable switches will understand BPDU(A)s.

It is recommended that capabilities exist to disable the RSTR mode on a per-port basis. When this option is selected the STP BPDUs are sent without ASPC flag set and the switch can interoperate

with other vendor switches on the RSTR-disabled port. BPDU(A) is sent periodically on RSTR enabled ports as previously defined.

The ASPC flag has the following advantages:

- ASPC should update non-core switch priority when switch receives BPDU with ASPC flag.
- Allows an unambiguous start of the automatic configuration.
- Allows for faster STP stabilization for non-RSTR core switches.

# of octets	ASPC BPDU				
2	protocol identifier				
1	**version = 1000 0001B**				
1	message type				
1	TCA	reserved	**ASPC**	TC	Flags
8	root ID				
4	cost of path to root				
8	bridge ID				
2	port ID				
2	message age				
2	max age				
2	hello time				
2	forward delay				

Figure 4-4 BPDU(A) Format

4.8 AUTOMATIC PORT ROLE SELECTION

The ASPC enables the automatic **port role** selection - upstream, downstream, peer and core ports can be identified. Observing the STP Bridge priority in incoming BPDU(A)s and comparing it to the current STP Bridge priority derived via ASPC allows this identification.

It is recommended that port role selection should follow performance of the STP Bridge Priority Selection Algorithm.

The following rules govern the port role selection:

- By default all ports should be initially configured as downstream ports.
- A port should change its role only if it is in the RSTR_ACTIVE state.
- Port's role should be verified after each received BPDU(A) where the neighbor switch's STP Bridge priority is different from previously observed STP Bridge priority in BPDU(A)s on this port.

- **Upstream Port** – A Port is classified as an upstream port when the neighbor switch's STP Bridge priority taken from BPDU(A) received on this port is smaller than the current STP Bridge priority. The upstream ports should exist only on non-core switches.
- **Peer Port** – A port is classified as a peer port when the neighbor switch's STP Bridge priority taken from BPDU(A) received on this port is equal to the current STP Bridge priority.
- **Core Port** – A port is classified as a core port when the neighbor switch's STP Bridge priority taken from BPDU(A) received on this port is equal to the current STP Bridge switch priority and equals zero.
- **Downstream Port** – A port not classified as upstream, peer or core is implicitly classified as a downstream port. This implies that the neighbor switch's STP Bridge priority, if any, received on this port is greater than the current STP Bridge switch priority.

A situation where the Priority and the Cost hierarchies are opposite may create downstream root ports. Figure 4-5 shows an example. The links from switch E are all at a very low cost (i.e. they have high bandwidth) compared to the other links. This creates to possibility that switch D selects one of its downstream ports as the root port if the cost D-F-E-A is less than the cost upstream from D to A.

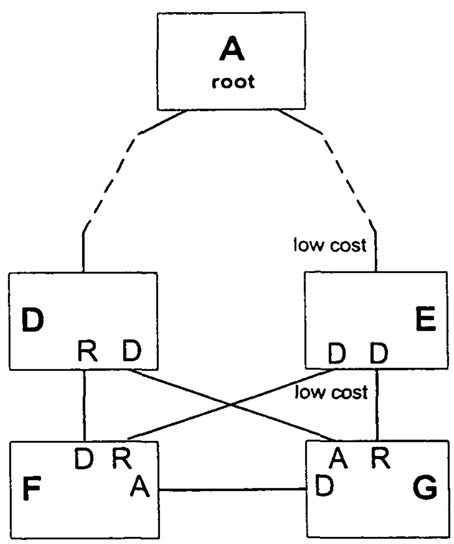


Figure 4-5 Downstream Root Port Scenario

This may create a potential loop if downstream links from switch E simultaneously fail: Switch G and F make a Rapid switchover to their alternate ports and a loop: G-D-F-G is created.

To avoid this possibility we have the following recommendation:

- When a switch wants to make a downstream port a root port or discovers that its root port is downstream, the switch disables transmission of BPDU(A) on all ports. All other functionality

of the switch is unchanged. If the root port later becomes an upstream or peer port, transmission of BPDU(A)s is re-enabled.

4.9 DISABLING STP ON A SINGLE SWITCH

A prerequisite for RSTR is STP. An RSTR_ACTIVE port should be connected to a neighbor switch that has enabled RSTR and STP.

Let's consider the example shown below. All switches have RSTR and STP enabled and full-duplex links interconnect switches. The switches receive BPDUs from one switch on each port only. Now let's assume that someone has disabled STP on switch C. This may be due to an improper switch configuration, or a transient condition. As a result switch C forwards each BPDU as a normal frame. Because of this forwarding condition, switches A, B and D receives BPDUs from two different switches on ports directly connected to switch C. It may be an undesirable state for a RSTR switch/network since the automatic port role classification could be unstable. As a result, the RSTR_ACTIVE port, which receives BPDUs from different switches, should be suspended (i.e. it should enter the RSTR_SUSPENDED state).

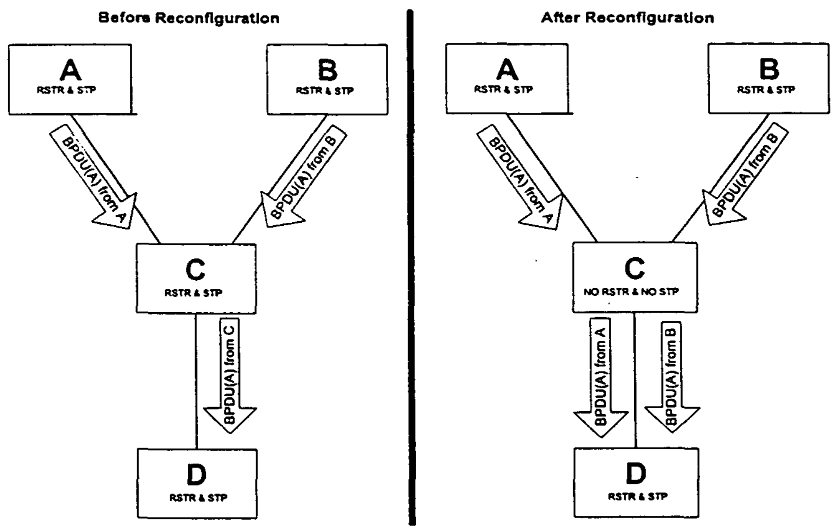


Figure 4-6 Disabling STP on switch

4.10 SWITCH POWER UP SEQUENCE

During the switch power up sequence all RSTR ports should follow the standard STP behavior in order to determine the Spanning Tree topology and to allow sufficient time to propagate the STP Bridge priority through the network. To provide this behavior, it is recommended that the Spanning Tree Protocol and RSTR state determination should run in parallel.

It is recommended that subsequent changes to the topology, like activation of a port, follow the modified STP behavior described in the next section.

5 RSTR PORT GROUP MAINTENANCE

5.1 RSTR PORT GROUP CREATION

The Automatic Switch Priority Configuration and the automatic port role selection help to enable the automatic **RSTR Port Group RPG** creation. A two-step procedure is recommended for performing RSTR Port Group creation:

- First RSTR Port Group port candidates should be found. This operation is done after ASPC priority settings and port role classification. The RSTR Group is created from upstream and peer ports that should comply with the following conditions:
  - RPG contains upstream and peer ports provided they are the root or blocking ports.
  - The blocking ports are the backup ports.
  - All RPG members are RSTR_ACTIVE.

- After selecting RPG member ports it should be determined if RPG is ready to perform the RSTR rapid switchover operation. RPG is ready to perform the rapid switchover when:
  - It contains at least two ports.

In general, additional timing considerations are not required due to the execution of ASPC on top of STP. This allows switch priorities to be learnt before all ports of the Spanning Tree Group move to either STP_FORWARDING or STP_BLOCKING state.

In general, as long as RPG has not been created, the standard STP recalculation is used in case of link failures.

5.2 RSTR PORT GROUP DELETION

The following events can delete RPG:

- Operator changes the switch configuration:
  - Disable STP or RSTR Mode on the switch.
  - Disable RSTR mode on all but one upstream port.
- STP recalculates topology:
  - STP sets two upstream ports in the STP_FORWARDING state after the topology recalculation.
  - STP sets at least one port in the STP_LISTENING or STP_LEARNING state.
- Link failures:
  - Only one or none upstream port stays active after a failure of another upstream port.
- Upstream port selection change:
  - Automatic port selection procedure changes role classification of existing upstream port(s) to downstream.

RPG is not affected when the switch receives the STP Topology Change Notification frame.

5.3 ADDING NEW UPSTREAM/PEER PORT TO RPG

There are several situations when new upstream port in RSTR mode may be added to an existing RPG. For example:

- The operator adds a new redundant link to an upstream switch, or a previously failed link again becomes operational
- The operator enables a logically disabled port connecting to an upstream switch
- The operator enables the RSTR mode on a port connecting to an upstream switch where the RSTR mode previously has been disabled

The discussion in this section applies as well to peer ports.

RPG may contain only one port in the STP_FORWARDING state. This port is the root port because it has received the best BPDU (wherein 'best' may be a implementatin-defined subjective qualification that may consider path cost, STP Bridge priority, port Id and/or other considerations). Two cases should be considered when a new port becomes RSTR_ACTIVE and becomes a member of RPG:

- Case 1:
  - The BPDU received on new port is "worse" then received on the current root port.
  - In this case STP sets the new port to the STP_BLOCKING state. This port is added to RPG as a backup port. No changes are needed to the forwarding path.
- Case 2:
  - The received BPDU received on new port is "better" then received on the current root port (wherein 'better' may be determined by implementation-defined considerations such as link speeds). Traffic should be forwarded through a better path. There is no sense, for example, to forward traffic through 10Mbps link and keep 100Mbps link blockedFigure 5-1 shows such situation. The link between switch A and C has failed and becomes active again. This link will, in general, have a better path to the root than the link between switches B and C in such configuration.

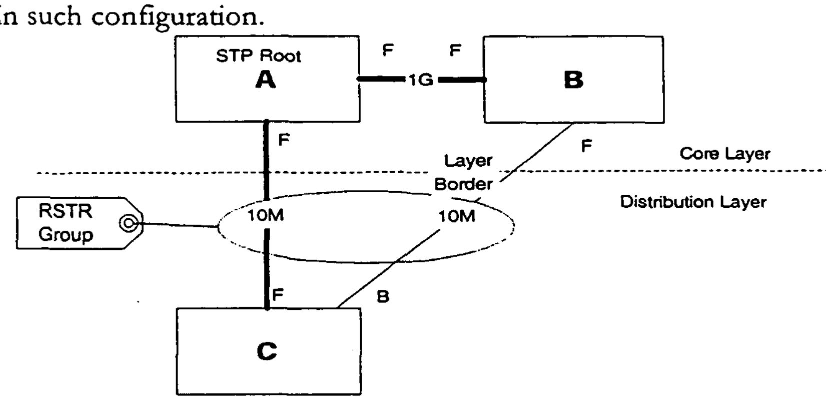


Figure 5-1 Root Port is active again

Switching to new port may be more complex than the case of rapid switchover when the root port fails due to the following issues:

- There is no guarantee that the other side of the new link is already in the STP_FORWARDING state (see the diagram below). The upstream switch C, which may have just been powered up, will perform standard STP recalculation and its downstream port will go through STP_LISTENING and STP_LEARNING states.

- When STP is notified that the new link is added then STP immediately sets the current root port into the STP_BLOCKING state and sessions are lost in such case.

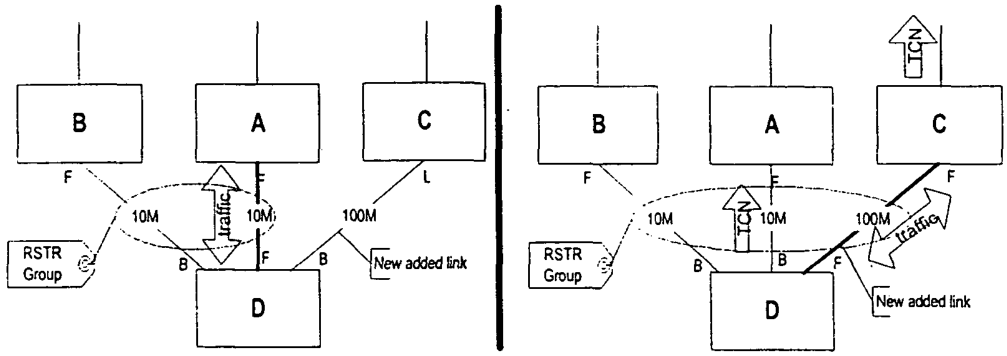


Figure 5-2 Adding New Link

In a 'standard' STP process, a new link on switch D would normally go through STP_LISTENING and STP_LEARNING states. However, this may lead to deletion of the RSTR Port Group. Therefore, in the RSTR process, it is recommended that a new port goes into the STP_BLOCKING state (c.f. section 4.4).

When a switchover from the current root port to new port is to occur, a delay may be used to help ensure that the other side of the new link is in the STP_FORWARDING state (other side of the link could go through the STP_LISTENING and STP_LEARNING states). The STP Forward Delay timer can trigger switching to new better port. It is recommended that the new port wait in the STP_BLOCKING state until the STP Forward Delay timer expires two times. This delay helps to ensure that ports on both sides of new link enter the STP_FORWARDING state at appropriate times.

When a switch adds a new STP_BLOCKING member port to RPG, it is recommended that the switch wait a delay period of, for example, twice the STP Forward Delay time (i.e. 30 seconds) before the new port can be used as a backup port. RSTR should maintain a sub-state of the STP_BLOCKING state:

- **RSTR_BLOCKING_TIMING** - a blocking port is waiting to become a full member of RPG, i.e. is subject to twice STP Forward Delay timer.
- **RSTR_BLOCKING_STANDBY** – a blocking port which is ready for use as a rapid switchover.

Note that when the new link is in the STP_BLOCKING state then all traffic should be forwarded through the current root port in order to protect all active sessions. RSTR should force

STP_FORWARDING state on the current root port in RPG against standard STP behavior that tries to block this port.

All addresses learnt on the old root port should be immediately aged. The TCN frame should be sent on the old root port, which just enters the STP_BLOCKING state. This may helps to speed up the address aging on other switches in the network. The upstream switch, which has new link connected, should also sends TCN frame (see Figure 5-2) (this is a standard STP behavior).

BPDUs, which the switch sends before the rapid switchover operation, contain the cost to the root through the old root port. After switchover BPDUs should contain the cost to the root through the new port.

Thus, this section defines changes to STP and switchover rules for new RPG backup ports:

- New upstream/peer port immediately goes into STP_BLOCKING state
- While the above change is enforced, the current root port remains in the STP_FORWARDING state
- If the new port has a better BPDU than the current root, a rapid switchover to the new port can be performed after 2 X STP Forward Delay timer.

5.4 CHANGING PATH COST

The change of a path cost on a port can be a result of several events:

- Operator reconfiguration of a link cost for a backup port in RPG
- Operator reconfiguration of a link cost somewhere upstream in the network changing the path cost on a backup port in RPG
- Rapid switchover upstream in the network – changing the path to the root results in a path cost change on ports in RPG

The last situation is discussed in section 6.2.

RPG should deal with all changes of path costs: RPG should always ensure that the root port is the port with the best path cost. If one of the backup ports becomes a 'better' root than the current root, then RPG should perform a rapid switchover to this port, subject to switchover rules defined in 6.5. (e.g. a root becoming a designated port prevents the switchover).

A difference between this situation and the previous one (adding new backup/peer port) is that the twice the STP Forward Delay timer delay may not be needed in this situation.

5.5 PEER LINKS

In a network with equal cost on all links, peer links never forward data traffic because STP always sets one end of the link in the STP_BLOCKING state on (see the link between switches C and D in Figure 5-3).

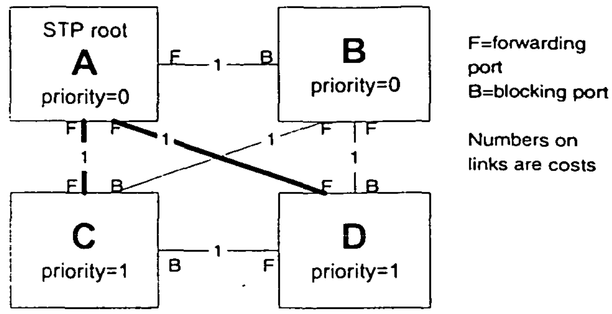


Figure 5-3. Peer port states in normal case

If different costs are assigned to different links, the switch's root port may be a peer port, and thus both ports on the peer link are STP_FORWARDING (see Figure 5-4).

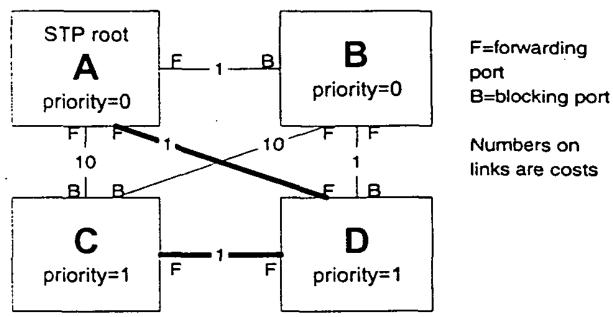


Figure 5-4. Peer port states in unusual case.

However, for the purpose of the rapid switchover, there is no problem with including peer ports that are either STP_BLOCKING or are STP_FORWARDING root ports in the RSTR group.

Adding a new peer port to RPG is equivalent to adding a new upstream port. In case the new peer port provides a better path to the root, this peer port is selected to be a root port and the procedure described in the previous section is applied.

5.6 REMOVING UPSTREAM PORT FROM RSTR PORT GROUP

A port should be removed from RPG when:

1. Port leaves the RSTR_ACTIVE state.
2. Port role classification is changed form upstream to downstream.
3. Operator disables RSTR mode or STP on a port.

Backup ports are simply removed from RPG. If it is the last backup port then RSTR Port Group is deleted too. When the root port is removed from RPG the situation is more complex and there are several cases to consider:

1. If the root port leaves the RSTR_ACTIVE state then RSTR performs rapid switchover.
2. If the root port role classification is changed from upstream to downstream or peer then STP should recalculate the topology and the RSTR Port Group is deleted.

5.7 FAST POLLING

To detect failures of cables, RSTR switches can apply the Link Aggregation Control Protocol which offers simple link status. The Link Aggregation Control exchanges information with a link partner on a regular basis and records information about a link partner state's and partner's knowledge.

6 RSTR RAPID SWITCHOVER

The general idea of RSTR Rapid switchover operation is shown in Figure 6-1. This figure shows only a part of a hierarchical network. There are two layers of switches. The switch from lower layer has two upstream links. One of the two upstream links is redundant. STP has blocked redundant links to prevent loops. As soon as the switch detects a failure of the upstream forwarding port or link, it immediately changes the state of the STP_BLOCKING port to the STP_FORWARDING State, without going via the STP_LISTENING and STP_LEARNING states.

The Automatic Switch Priority Configuration guarantees hierarchical network configuration with a proper setting of the STP Bridge priority (= network layer) and port role selection. The hierarchical network structure ensures that the RSTR Rapid switchover operation will work properly. The network hierarchy makes certain that all downstream ports are in the STP_FORWARDING state, and that only one upstream port is in the STP_FORWARDING state. This helps to ensure that the downstream switch will not switch to a link that is in the STP_BLOCKING state (as seen from the upstream switch's point of view) see Figure 6-1.

In general, only a downstream switch can perform a rapid switchover, i.e. the RSTR core switches cannot perform the rapid switchover.

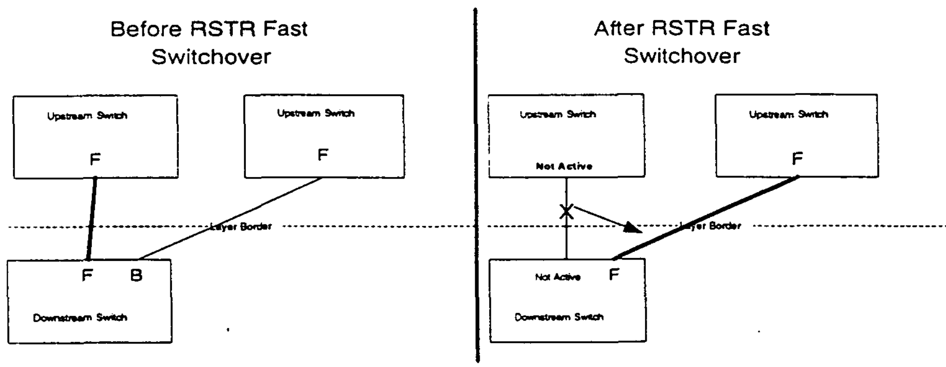


Figure 6-1 RSTR Rapid switchover

6.1 TRIGGERING RSTR RAPID SWITCHOVER

The RSTR Port Group creation enables the rapid switchover operation. The failure of the root port in the RSTR Port Group triggers the RSTR rapid switchover.

An RSTR Port Group may have more than one port in STP_BLOCKING state. During the switchover operation, RSTR should select one of the backup ports and change its state to STP_FORWARDING. The backup port selection criteria is as follows:

- First, the port that has the best path to root is selected.
- If there are more than one backup port which have the same costs to root, standard spanning tree procedures are used to determine the "best" port.

6.2 RSTR RAPID SWITCHOVER AND STP COSTS

RSTR rapid switchover operation on one switch can have influence on RSTR Port Group configuration in other parts of the network.

The example in Figure 6-2 explains this issue. All links in this network are Fast Ethernet. The numbers on the links show their costs. Switches C and E create RSTR Port Groups. Port 1 is selected as root port on switch C because it has better cost to the root than port 2. For switch E, the costs to the root on port 1 and 2 are the same. Port 1 is selected as the root port because it has a lower port number.

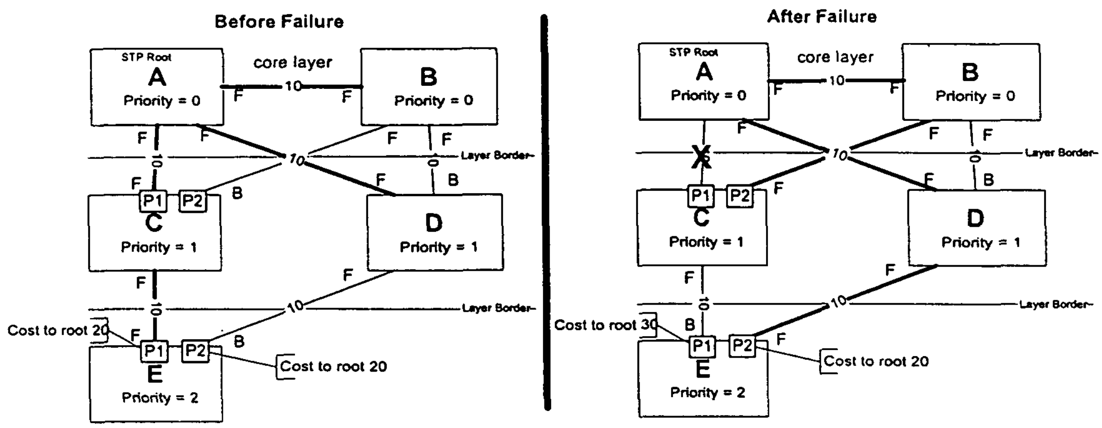


Figure 6-2 RSTR Switchover and STP Costs

As an example, assume that the link between switches A and C fails. Switch C performs the rapid switchover operation and port 2 is now the root port. Because of this, the root port (port 1) on switch E has a worse path cost to the root than the backup port (port 2). In such a case, STP expires the saved BPDU before accepting the worse BPDUs from E. STP then puts port 1 in the STP_BLOCKING state and sets port 2 to the STP_LISTENING state. As a result current sessions are lost. RSTR should change standard STP behavior in order to protect sessions.

To reduce or solve this problem, the standard spanning tree algorithm (which replaces old information with new information when the new information is better) is modified. The modified algorithm incorporate the following behaviors:

- Old information is replaced with new information when the new information is bettor or when it comes from the same bridge (MAC address) and port as the old information.
- The hello timer is started if the bridge has become the root bridge.

RSTR detects backup ports that have better path cost to the root than the current root port by observing incoming BPDUs. A backup port with a better path cost to the root is set in the STP_FORWARDING state and becomes the new root port (provided it is in the RSTR_BLOCKING_STANDBY state). The former root port is set in the STP_BLOCKING state and becomes a backup port.

All addresses learnt on the old root port are aged out immediately. TCN message should be sent on the old root port, which just enters STP_BLOCKING state. It forces address aging on other switches in the network.

6.3 ADDRESS DATABASE MANAGEMENT DURING RSTR SWITCHOVER

The RSTR rapid switchover dramatically reduces the STP convergence time but is not enough to prevent session loss. Switches affected by failures should update their address tables in order to continuously forward session data traffic to right destinations. This goal is achieved by the following operations:

The switch that performs the rapid switchover operation should immediately age all addresses known on the failing port (old root port).

- After the rapid switchover, the switch should send immediately TCN-frame (Topology Change Notification). This does not cause the spanning tree recalculation but it forces fast/immediate address aging.
- As an additional address table update measure, the switch, which detects a failure of one of its downstream forwarding port, should send a TCN-frame.
- After reception TCN frame switch in RSTR mode should immediately age addresses to reroute active sessions within a short time period (e.g., 3 seconds).

Let's consider an example shown in Figure 6-3 and Figure 6-4. Traffic between file server FS1 and workstation WS is forwarded throughout switches A, B and D. Switch A knows FS1 address on port 1 and WS address on port 2. Switch D knows FS1 address on port 1 and WS address on port 3. Ports 1 and 2 create RSTR Group on switch D.

Now, lets assume that link d has failed. Switch D has changed port 2 to the STP_FORWARDING state and this port now becomes root port. FS1 address was known on port 1 on switch D but after RSTR switchover this address is immediately aged and should be learned again by the normal flooding mechanism (on port 2).

If, at the moment of failure, traffic was going from WS to FS1 then switch A should now receive frames from WS on port 3. As a result switch D correctly forwards all frames from WS to FS1 along the new path (Switch D, C, A, FS1).

The situation is more complex if, at the moment of failure, traffic was going from FS1 to WS. Switch A knows WS address on port 2 and frames from FS1 to WS are forwarded to switch B, which does not have path to WS. To prevent session loss, fast address aging on port 2 of switch A is recommended. This is accomplished by sending TCN-frame by switch B when failure of a port in this switch is detected. TCN-frame is a signal for immediate aging.

It may happen that file server FS2 has an active session with workstation WS. Switch C has known WS address on port 1 before the link failure. Traffic between FS2 and WS has been forwarded through switches C, A, B and D. This path is no longer valid and it should be replaced by a path through switches C and D. Switch C should age out WS address on port 1 and flood all frames to WS. In this case it is recommended that switch D sends TCN-frame on the new root port (port 2) – immediately after moving this port to the STP_FORWARDING state.

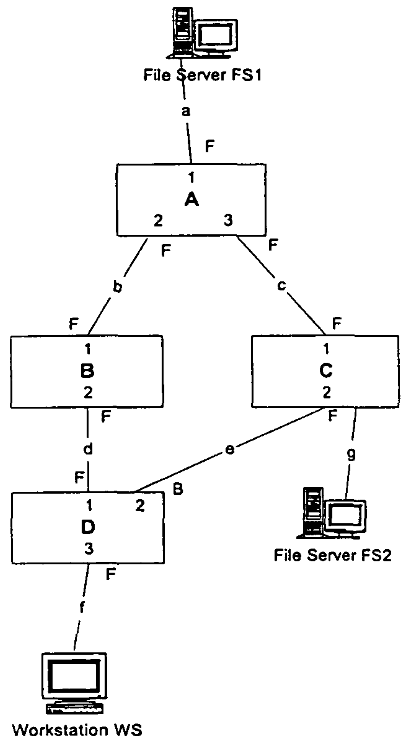


Figure 6-3 Before Switchover

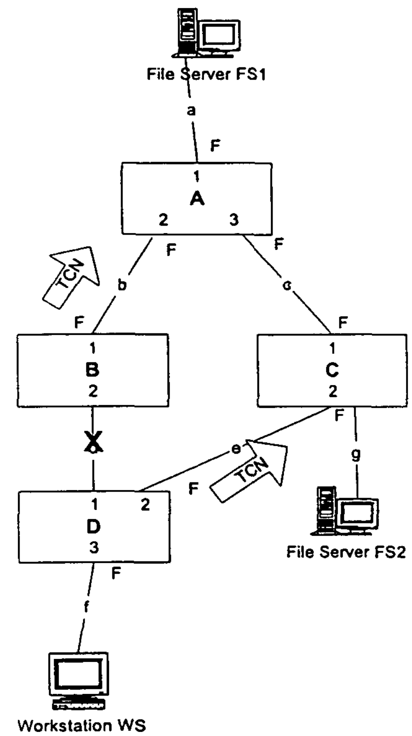


Figure 6-4 After Switchover

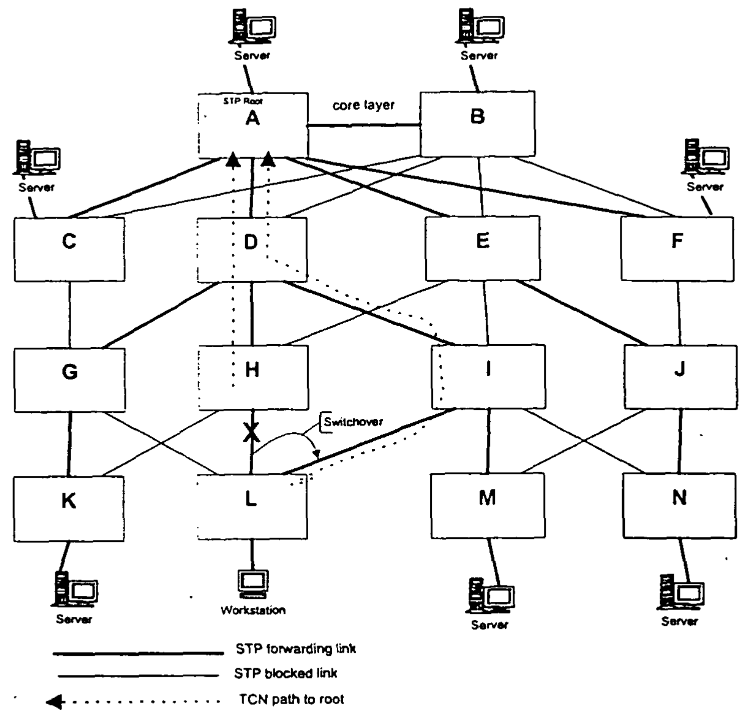


Figure 6-5 TCN paths to root I

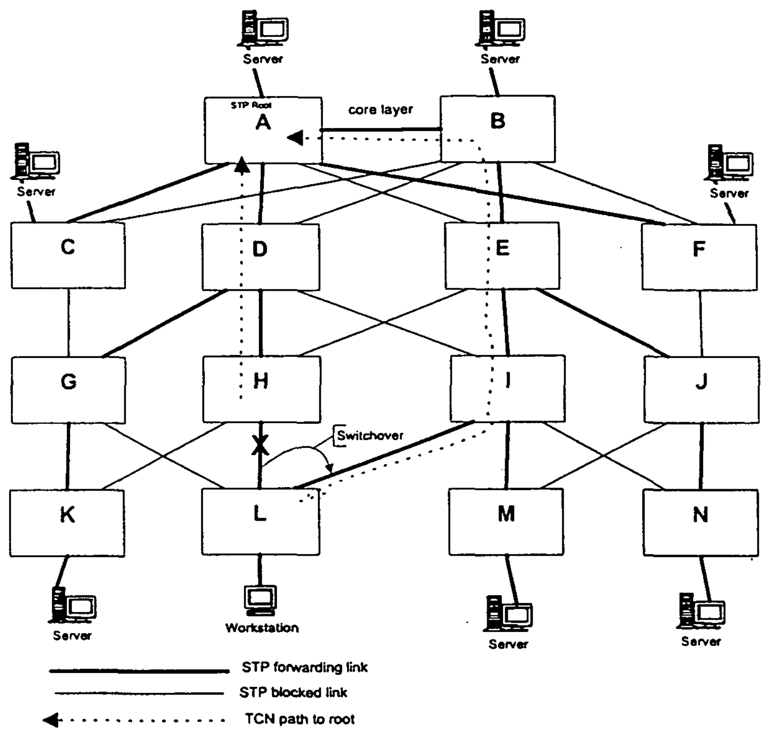


Figure 6-6 TCN Paths to Root II

Note that after the rapid switchover operation switches that are deployed on TCN frames paths to the root can choose wrong ports in forwarding process if address databases at these switches do not reflect the current network topology. Those switches are switches A and C inFigure 6-4, switches A, D and I in Figure 6-5 and switches A, B, D, E and I in Figure 6-6. To prevent this, each switch in the RSTR mode should immediately age out all addresses upon reception of TCN-frames.

This could cause flooding after the rapid switchover but the immediate aging operation allows rapid rerouting of active sessions(e.g., within 3 second).

Other switches in the network may not need to update theirs Address Databases and may follow standard STP rules in their address aging procedure.

6.4 LINK FAILURE AND TRANSMIT, RECEIVE QUEUES

When a port becomes inactive, frames in transmit and receive queues should be flushed in order to decrease probability of improper frames (packet) sequencing.

The frames in transmit queue should be flushed because these frames may be obsolete when this port becomes active again and they can cause improper packet sequencing. These packets may have been retransmitted through redundant path after the rapid switchover operation.

6.5 RAPID SWITCHOVER RULE SUMMARY

An RSTR groups consists of a number of upstream and peer ports. In general, one port in the RSTR group is STP_FORWARDING (root port), the other ports are STP_BLOCKING. Some of the STP_BLOCKING ports are RSTR_BLOCKING_STANDBY, while other ports are RSTR_BLOCKING_TIMING.

The following general rule applies:

- Rapid switchover should not occur as the result of the root port becoming a designated port.

With this provisions, the following rules govern rapid switchover:

If the STP_FORWARDING port goes down or is disabled, then...
...if there are any RSTR_BLOCKING_STANDBY ports, a rapid switchover to the best such port is performed (section 6.1 describes the port to choose),
...otherwise, standard STP procedures are followed.

If a RSTR_BLOCKING_STANDBY port hears a better cost than the STP_FORWARDING port, then...
...a rapid switchover to the RSTR_BLOCKING_STANDBY port is performed.

If a RSTR_BLOCKING_TIMING port hears a better cost than the STP_FORWARDING port, then...
...the STP_FORWARDING port remains STP_FORWARDING until the RSTR_BLOCKING_TIMING port is promoted to RSTR_BLOCKING_STANDBY; then (if the required conditions are still present) a rapid switchover is performed.

If a STP_BLOCKING port goes down or is disabled, it is simply removed from RPG; no special action is required.

If a STP_BLOCKING port becomes a designated port, it is removed from RPG and goes through STP_LISTENING and STP_LEARNING states.

Two special cases of the above rules are worth mentioning here:

- Assume RPG consists of three ports P_r, P_t, and P_s. Further assume that P_r is the root port, P_t is RSTR_BLOCKING_TIMING, and P_s is RSTR_BLOCKING_STANDBY. Now, if P_t hears a better cost than the forwarding port, P_r remains forwarding while we wait for P_t to be promoted to RSTR_BLOCKING_STANDBY. If P_r fails during this waiting time, we will do a rapid switchover to P_s, and then – when P_t becomes RSTR_BLOCKING_STANDBY – we will do another rapid switchover to P_t.

- A special case of the "better cost heard"-cases occurs when a new root is heard. A rapid switchover may also be performed in this case.

The above Rapid Switchover rule which states, that the best RSTR_BLOCKING_STANDBY port should be selected as the target port for a switchover, even if a better RSTR_BLOCKING_TIMING exists (i.e. to skip better RSTR_BLOCKING_TIMING ports), complicates the implementation considerably. Hence, implementations may simply ignore worse RSTR_BLOCKING_STANDBY ports when a better RSTR_BLOCKING_TIMING port exist. This disallows Rapid Switchover in a twice Forward Delay timer period after a new and better (upstream) link is added even if there should exist RSTR_BLOCKING_STANDY ports.